ISSUES OF VALIDITY: WHAT CAN BE LEARNED FROM THE STUDENT EVALUATION OF TEACHING

Dennis E. Clayson, College of Business, University of Northern Iowa, Cedar Falls, Iowa 50614-0126, (319) 273-6015, dennis.clayson@uni.edu

ABSTRACT

The student evaluation of teaching (SET) can be seen as a type of job performance measure, and as such can act as an example, not only of academic performance, but also of attempts by management and HR specialists to measure non-objective performance. Since SET has been extensively studied, the pitfalls, problems, and applications of these evaluations can act as guides for anyone attempting to measure performance.

When looking at the validity of an instrument, three questions need to be addressed. 1) How will the instrument be utilized? Or, what is the ultimate purpose or reason for using the instrument? 2) Does the instrument meet the demands of research methodology and logic? And/or, what type of validity is required to demonstrate these demands? And 3) Does it matter?

How will the instrument be utilized?

One issue that needs to be addressed early in any discussion of validity is the use to which an instrument may be applied. SET instruments, as any other performance measure, may have a number of applications. They could be utilized for: 1) instructional improvement, 2) evaluation of performance with personnel and managerial implications, and/or 3) necessary feedback to comply with legislative, administrative, or student demands. Validity in one utilization does not guarantee validity in another.

Does the instrument meet the demands of research and logic?

SET is seldom controlled for contaminating factors. One area of concern is the degree to which students project unrelated factors onto the instructor or on the class itself. Another problem is the creation of scales. Whenever a question is asked that can be quantified, a scale is assumed by the type of answer given, or which *can* be given. The appropriateness of a statistical analysis is dependent upon this scale The problems created by scales in evaluations are often overlooked. Nominal and ordinal scales are common on the instruments. Administrators must be careful not to make statistical and qualitative comparisons that are not justified by their forms.

What type of validity is required?

Just as a person can be said to have a certain personality structure based on the results of paper and pencil tests, instructors are said to be a “good” or “bad” teacher based on the evaluations. Since hypothetical constructs, such as “effective” teaching, have no form or dimension that can be objectively measured, a number of different associations must be established. This raises the possibility that an instrument may be valid in certain ways and simultaneously invalid in others.

Face validity exists when an instrument appears to be measuring what the respondent thinks it should be measuring. It is related to the respondent's experience with the instrument. Since SET instruments are created and sanctioned by the institution, ask questions generally associated with instruction, and are administrated in formalized manners, it would be assumed that SET has face validity.

Content validity is said to exist if the questions on an instrument can be logically said to cover the domain of the construct that the instrument is intended to measure. Since most institutions do not have a clearly defined definition of what they are attempting to measure, the content validity of evaluations is suspect. *Concurrent* and *predictive validity* in SET is considered to be strong to adequate. Results are related to other measures of teaching effectiveness and seem to be predictive of measures made by former students, self-reports, and of trained observers.

In terms of *construct validity*, the evaluations have been found to have *convergent validity*, but to be lacking in *discriminant*, and *divergent validity*. This relationship would be expected if the evaluations were measuring one, or a few, of the dimensions deemed relevant to teaching, but not all. A number of attempts have been made in business education to investigate *nomological validity* with mixed success. There are two difficulties. First, they typically establish a network from students' perceptions of teaching either by gathering primary data and/or by looking at SET instruments that already exist. It could be argued that instead of finding A is related to B, the procedures demonstrates that A is related to A. In addition, what is "good" or "effective" teaching has seldom be defined.

It is possible for an instrument to have *utilitarian validity* when lacking almost all else. An instrument could hypothetically, be useful as a tool to achieve an end irrespective of any validity to related theoretical constructs. An extreme example would a politician using an invalid poll to win an election. More commonly, an evaluation may be given simply because it is required to satisfy some tangential or even unrelated regulation or goal, and not because anyone is directly influenced by specific results of the measurement.

Does it matter?

As this review shows, the validity of an instrument is complicated by a number of factors, but in the case of SET, the instruments continue to be almost universally utilized. Is this a problem? Most of the defenders of the SET system come from the colleges of education and their strong defense is driven by both research and ideological concerns. However, much of the current research does not support such optimism. Nevertheless, when organizations and, in some cases, complete societal units, demand performance assessments, the validity of instruments making those assessments may be secondary to the ideological or societal reasons for the demand. Utilitarian validity can be maintained by organizations, even in the face of potential lawsuits, by careful wording of the consequences of the evaluations.

REFERENCES AVAILABLE UPON REQUEST